AIRI at RRG24: LLaVa with specialised encoder and decoder

V. Samokhin[1], **M. Munkhoeva**[1], **D. Umerenkov**[1]*, **E. Kuzmina**[1,2], **I. Oseledets**[1,2], **D. Dylov**[1,2]
[1]AIRI, [2] Skoltech
{samokhin,munkhoeva,umerenkov,kuzmina,oseledets,dylov}@airi.net

Abstract

We present a new approach to generating the "Findings" and "Impression" sections in the chest X-rays radiology reports, developed as part of the shared radiology task at BioNLP 2024. By integrating a DINOv2 vision encoder trained on medical data with specialized biomedical large language model using the LLaVA framework, our method addresses complex medical semantics and diverse findings in imaging. We use datasets from PadChest, BIMCV-COVID19, CheXpert, OpenI, and MIMIC-CXR. The evaluation metrics demonstrate our method's effectiveness and the potential for automating the generation of radiology reports.

1 Introduction

The automatic generation of radiology reports from chest X-rays is a challenging and significant task in the field of biomedical natural language processing (BioNLP). The growing volume of medical imaging data and the limited number of radiologists necessitate the development of robust automated systems to assist in report generation. Such systems not only have the potential to improve clinical workflow efficiency but also to ensure consistency and comprehensiveness in radiological interpretations.

In recent years, advancements in deep learning and natural language processing have paved the way for innovative approaches to tackle this task. The new approaches typically involve the integration of convolutional neural networks (CNNs) or visual transformers for image feature extraction with recurrent neural networks (RNNs) or transformers for text generation (Selivanov et al., 2023). Despite the progress, challenges such as capturing complex medical semantics, handling diverse imaging findings, and ensuring the clinical accuracy of generated reports remain.

This paper explores a new method for generating the Findings and Impression sections of radiology reports from chest X-rays. Our approach is to combine a vision encoder, self-trained on medical data, with specialized biomedical LLM for text generation, using LLaVA framework. This work was done as a part of Radiology Report Generation ahared task at BioNLP 2024 Workshop (Xu et al., 2024) using the data provided by the organizers. The metrics were calculated using the ViLMedic platform (Delbrouck et al., 2022b).

2 Data

2.1 Training and validation data

The data from 5 datasets where combined to create the competition training and validation datasets: PadChest(Bustos et al., 2020), BIMCV-COVID19(Vayá et al., 2020), CheXpert(Chambon et al., 2024), OpenI(Demner-Fushman et al., 2012) and MIMIC-CXR(Johnson et al., 2019). The training and validation sets are grouped by study but not by subjects. The official language of PadChest and BIMCV-COVID19 is Spanish, and their reports have been translated using GPT-4 by the shared task organizers.

The data consists of radiology studies, each containing one or more chest X-ray images in various projections. Each study also includes Impression and Finding texts. Some studies have only the Impression or only the Findings section, while others have both.

2.2 Testing Data

The studies in the test sets are unseen studies provided by organizers. Public test sets for impression and findings contain both study images and ground truth texts while private test set contains only images.

*V.S., M.M., and D.U. contributed equally.

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 603–607
August 16, 2024.

Dataset	Findings	Impressions
PadChest	101,752	-
BIMCV-COVID19	45,525	-
CheXpert	45,491	181,619
OpenI	3,252	3,628
MIMIC-CXR	148,374	181,166
Total	**344,394**	**366,413**

Table 1: Training dataset statistics.

Dataset	Findings	Impressions
CheXpert	1,112	4,589
BIMCV-COVID19	1,202	-
PadChest	2,641	-
OpenI	85	92
MIMIC-CXR	3,799	4,650
Total	**8,839**	**9,331**

Table 2: Validation dataset statistics.

Dataset	Findings	Impressions
public test-set	2,692	2,967
hidden test-set	1,063	1,428

Table 3: Testing datasets statistics

2.3 Data preprocessing

Due to technical limitations, we only used the first two images from each study. Studies with only one image were not further processed. For studies with more than one image, the first two images were stitched together horizontally. No additional preprocessing was applied to the texts.

3 Evaluation metrics

In the evaluation of radiology report summarization systems, several metrics are commonly used to assess the performance and accuracy of the generated summaries. These metrics ensure that the summaries produced by the models are not only syntactically and semantically correct but also factually accurate. The metrics used in this competition where BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), BERTScore (Zhang et al., 2019), F1-CheXbert (Smit et al., 2020), and F1-RadGraph (Delbrouck et al., 2022a).

3.1 BLEU (Bilingual Evaluation Understudy)

BLEU-4: This metric is widely used for evaluating machine translation systems. It measures the precision of n-grams in the generated summary by comparing it to one or more reference summaries. BLEU-4 specifically considers 4-gram overlaps, providing a robust measure of how many 4-grams in the generated text appear in the reference texts. However, it does not account for recall or the contextual meaning of words.

3.2 ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

ROUGE-L: ROUGE is predominantly used for evaluating automatic text summarization. ROUGE-L measures the longest common subsequence (LCS) between the generated summary and the reference summary. This metric emphasizes recall by capturing the longest sequence of words that appear in both the generated and reference summaries, thus reflecting the ability of the summary to include important information.

3.3 BERTScore

BERTScore: This metric computes the similarity between the generated and reference texts using pre-trained BERT embeddings. It calculates a similarity score for each token in the candidate sentence with each token in the reference sentence. BERTScore accounts for the semantic meaning of words, making it more robust against synonyms and paraphrasing compared to BLEU and ROUGE.

3.4 F1-CheXbert

F1-CheXbert: This metric evaluates the factual correctness of the generated summaries with a focus on specific medical conditions mentioned in radiology reports. CheXbert is a specialized tool designed to extract medical observations from radiology reports. The F1 score is calculated based on the precision and recall of these extracted observations, ensuring that the generated summaries accurately reflect the medical conditions described in the reference summaries.

3.5 F1-RadGraph

F1-RadGraph: Similar to F1-CheXbert, this metric evaluates the factual correctness of the summaries using the RadGraph dataset. RadGraph focuses on extracting entities and the relations between them from radiology reports. The F1-RadGraph score measures the accuracy of these extractions, comparing the generated summaries to the reference summaries to ensure that the critical entities and their relationships are accurately captured.

These metrics collectively provide a comprehensive evaluation framework for radiology report summarization systems. BLEU and ROUGE focus on the surface-level n-gram overlaps, while BERTScore provides a deeper semantic evaluation. F1-CheXbert and F1-RadGraph ensure the factual accuracy of medical details, which is crucial for clinical applications.

4 Methods and Results

We used the LLaVA model (Liu et al., 2024) with a DINOv2 encoder (Oquab et al., 2023) and OpenBio-LLM-8B (Ankit Pal, 2024) as a text decoder. The whole pipeline was implemented using HuggingFace's *transformers* (Wolf et al., 2020) and *trl* (von Werra et al., 2020) libraries.

For image encoding we used a DINOv2 Model with the following parameters:

- **Model**: ViT-base 14, initialized from torch.hub's `dinov2_vitb14`
- **Patch size**: 14
- **Number of parameters**: 86M
- **Time and Resources:** 4xA100 80GB GPUs, Training Total Time: 2 days
- **Dataset:** MIMIC-CXR Train, downsampled to 518 px
- **Batch size per GPU: 50**
- **Base Learning Rate:** 0.001

For text generation, we used OpenBioLLM-8B, an open-source language model designed specifically for the biomedical domain.

- **Training type:** LoRA on LLM's Attention matrices (r=64, alpha=16) + MM projector
- **Architecture:** OpenBio-LLM-8B + in-house DINOv2 trained on MIMIC-CXR Train
- **Time and Resources:** 5 epochs, 8xH100 80GB GPUs, DeepSpeed Zero-3; Training Total Time: 2 days
- **Batch size per GPU:** 8, gradient accumulation: 2
- **Base Learning Rate:** 0.001, cosine schedule, warmup: 0.15
- **Optimizer:** Adam

Vanilla approach to fine-tune LLaVA model with language model unfreezed resulted in rapid overfitting, thus we opted for PEFT methods (Mangrulkar et al., 2022), namely LoRA (Hu et al., 2022).

We used the same model for generating both impression and findings, using different prompts: either "Write findings for this X-ray." or "Write impression for this X-ray.".

We used the following system prompt, inspired by LLaVA-Med (Li et al., 2024):"You are a large language and vision assistant. You are designed to assist human with a variety of medical visual content and clinical research tasks using natural language. Follow the instructions carefully and provide clinically valid answers."

Our results on hidden test sets are presented in Table 4 and Table 5.

Table 4: Findings - hidden test set (1063 samples)

Metric	e-health csiro	maira	airi
BLEU4	11.68	11.24	9.97
ROUGEL	26.16	26.58	25.82
Bertscore	53.80	54.22	52.42
F1-cheXbert	57.49	57.87	54.25
F1-RadGraph	28.67	25.48	25.29

Table 5: Impressions - hidden test set (1428 samples)

Metric	e-health csiro	maira	airi
BLEU4	12.33	11.66	10.91
ROUGEL	28.32	28.48	27.46
Bertscore	50.94	51.62	49.55
F1-cheXbert	56.97	53.27	52.32
F1-RadGraph	27.83	25.26	24.67

Our relatively simple model demonstrates strong performance in generating radiology reports. We attribute this success to the use of a specialized image encoder and a specialized large language model. Future improvements can be realized by employing larger models and fully using the available image data, which would likely enhance the competition metrics of the generated reports.

References

Malaikannan Sankarasubbu Ankit Pal. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. 2020. Padchest: A large chest x-ray image dataset with multilabel annotated reports. *Medical image analysis*, 66:101797.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P. Langlotz. 2024. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022a. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360.

Jean-benoit Delbrouck, Khaled Saab, Maya Varma, Sabri Eyuboglu, Pierre Chambon, Jared Dunnmon, Juan Zambrano, Akshay Chaudhari, and Curtis Langlotz. 2022b. Vilmedic: a framework for research at the intersection of vision and language in medical ai. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 23–34.

Dina Demner-Fushman, Sameer Antani, Matthew Simpson, and George R Thoma. 2012. Design and development of a multimodal biomedical information retrieval system. *Journal of Computing Science and Engineering*, 6(2):168–177.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alexander Selivanov, Oleg Y Rogov, Daniil Chesakov, Artem Shelmanov, Irina Fedulova, and Dmitry V Dylov. 2023. Medical image captioning via generative pretrained transformers. *Scientific Reports*, 13(1):4171.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. 2020. Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519.

Maria De La Iglesia Vayá, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, Francisco García-García, et al. 2020. Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients. *arXiv preprint arXiv:2006.01174*.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J. Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, Curtis Langlotz, and Jean-Benoit Delbrouck. 2024. Overview of the first shared task on clinical text generation: Rrg24 and "discharge me!". In *The 23rd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Bangkok, Thailand, August. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.